APPLICATION OF EM ALGORITHM ON MISSING CATEGORICAL DATA ANALYSIS

NORAINI BINTI HASAN

A report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Science (Mathematic)

Faculty of Science

Universiti Teknologi Malaysia

DECEMBER 2009

To my beloved husband, son and all my family members

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my thesis supervisor, Assoc. Prof. Dr. Ismail b. Mohamad, for encouragement, guidance, critics and friendship. Without their continued support and interest, this thesis would have never been the same as presented here. Librarians at UTM also deserve my special thanks for their assistance in supplying the relevant literatures.

My colleagues should also be recognised for their support and the assistance provided at various occasions. Their views and tips are useful indeed. My sincere appreciation also extends to my beloved husband and son, my family and also not forgotten my in-laws, for their understanding and sacrificial. Unfortunately, it is not possible to list all of them in this limited space.

ABSTRAK

Algoritma EM merupakan salah satu daripada kaedah untuk menyelesaikan masalah berkaitan dengan data tidak lengkap berdasarkan kepada satu rangka lengkap. Algorithma EM merupakan satu pendekatan parametrik untuk mencari taksiran ML data tidak lengkap. Algorithma ini terbahagi kepada dua langkah, dimana langkah pertamanya, langkah Ekspektasi, atau lebih dikenali sebagai langkah E, mencari ekspektasi kepada loglikelihood, bersyarat kepada data yang dapat diperolehi dan anggaran terkini, . Langkah kedua, langkah Pemaksimuman atau langkah M dimana ia akan memaksimumkan nilai loglikelihood untuk mencari satu anggaran parameter yang baru. Prosedur ini berlaku berselang- seli antara kedua-dua langkah ini sehingga anggaran parameter tersebut malar.

ABSTRACT

Expectation- Maximization algorithm, or in short, EM algorithm is one of the methodologies for solving incomplete data problems sequentially based on a complete framework. The EM algorithm is a parametric approach to find the Maximum Likelihood, ML parameter estimates for incomplete data. The algorithm consists of two steps. The first step is the Expectation step, better known as E-step, finds the expectation of the loglikelihood, conditional on the observed data and the current parameter estimates; say . The second step is the Maximization step, or M-step, which maximize the loglikelihood to find new estimates of the parameters. The procedure alternates between the two steps until the parameter estimates converge to some fixed values.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF SYMBOLS

	The observed value
	The missing value
	Number of observations or Total counts
	Estimates of
	Current estimates of
	The counts in cell (,)
	The observed value for
	The probability that an observation falls in cell (,)
( )	*r*th estimates of
	Observed frequencies
	Expected frequencies

CHAPTER 1

INTRODUCTION

1.1 PROBLEM STATEMENT

Incomplete table is referred to the table in which the entries or information on one or more of the categorical variables are missing, a prior zero or undetermined (Fienberg, 1980). Missing data treatment is an important data quality issue in data mining, data warehousing, and database management. Real-world data often has missing values.

The presence of missing values can cause serious problems when the data is used for reporting, information sharing, and decision support. First, data with missing values may provide biased information. For example, a survey question that is related to personal information will more likely be left unanswered for those who are more sensitive about privacy. Second, many data modeling and analysis techniques cannot deal with missing values and have to cast out a whole record value if one of the attribute values is missing. Third, even though some data modeling and analysis tools can handle missing values, there are often restrictions in the domain of missing values. For example, classification systems typically do not allow missing values in the class attribute.

Missing data always becomes the main obstacles for the researchers to further their studies. Some researcher will just ignore, truncate, censor, or collapse with those missing data. This might able to make the problem easier but it will lead to inappropriate conclusion and confusion. Therefore, a proper strategy should be used to treat such missing data.

1.2 OBJECTIVE OF THE STUDY

This research is carried out with some objectives as listed below:

1) To apply the EM algorithm on multinomial model in missing categorical data analysis.
2) To compare the results of independence test for complete and incomplete data.

1.3 SCOPE OF THE STUDY

This study is concentrated on the contingency table where some missing values are present and thus the EM algorithm will be applied on it. Only Missing At Random (MAR) data and Not Missing At Random (NMAR) data are considered in this study.

1.4 SIGNIFICANCE OF THE STUDY

The EM algorithm will be successful in dealing with missing data values in contingency table or in other words we can say that we can find the missing values by applying the EM algorithm. By the end of this study, we will discover a new dimension of problem such as the missingness mechanism which will have a direct impact or effect on the missing values.